

Comparative genomics tools for biological discovery

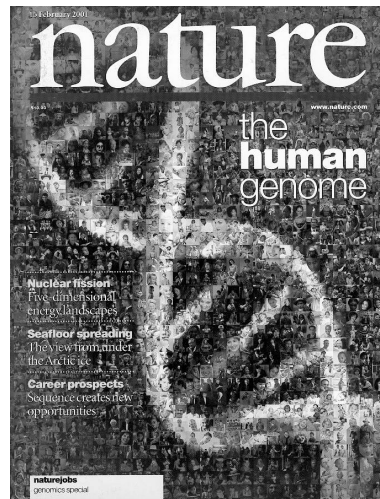
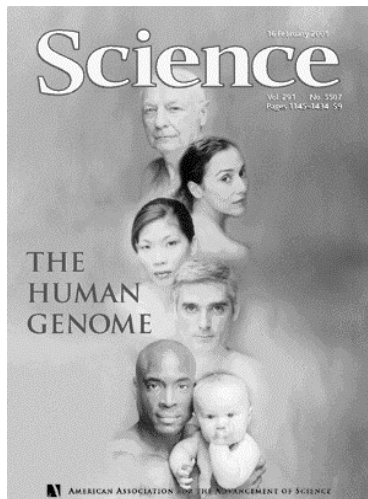
Inna Dubchak, Ph.D.

Staff scientist

Lawrence Berkeley National Laboratory

ildubchak@lbl.gov

The Human genome



From the Nature paper:

The next steps:

Developing the IGI (integrated gene index) and IPI (integrated protein index)

RefSeq: 14,200 Genscan: 47,440 Ensembl: 28,560

Large-scale identification of regulatory regions

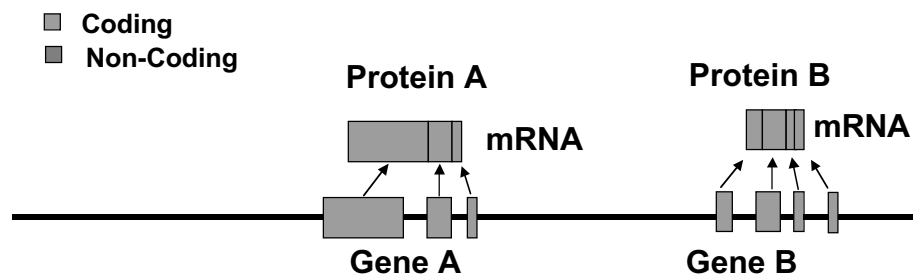
Sequencing of additional large genomes

Completing the catalogue of human variation

From sequence to function

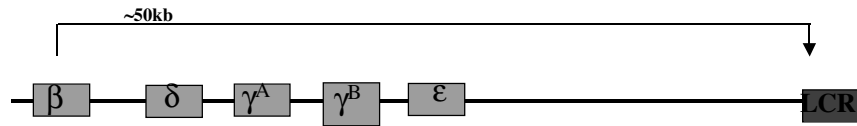
"

1-2% Coding



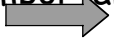
Distant Non-Coding Sequences Causing Disease

β-Thalassemia



Disease	Gene	Distance
Campomelic dysplasia	SOX9	850kb
Aniridia	PAX6	125kb
X-Linked Deafness	POU3F4	900kb
Saethre-Chotzen syndrome	TWIST	250kb
Rieger syndrome	PITX2	90kb
Split hand / split foot malformation	SHFM1	450kb

Background

- Evolution can help!
- In general, functionally important sequences are conserved  Conserved sequences are functionally important
- Raw sequence can help in finding biological function

Comparison of 1196 orthologous genes (Makalowski et al., 1996)

- Sequence identity:
 - exons: 84.6%
 - protein: 85.4%
 - introns: 35%
 - 5' UTRs: 67%
 - 3' UTRs: 69%
- 27 proteins were 100% identical

Integrating data into more powerful gene prediction
~~models than with human genomic sequence alone~~

Comparing sequences of different organisms



- Helps in gene predictions
- Helps in understanding evolution
- Conserved between species non-coding sequences are reliable guides to regulatory elements
- Differences between evolutionary closely related sequences help to discover gene functions

Challenges

- Sequence at different stages of completion, difficult to compare
-
- Whole genome shotgun → Partial Assemblies
- Finished BACs
- Fast and accurate analysis
- Scaling up to the size of whole genomes

<http://www-gsd.lbl.gov/vista>

VISTA
VISUALIZATION TOOLS FOR ALIGNMENTS

WELCOME to the homepage for VISTA, Visualization Tool for Alignments.

USE VISTA on the WEB
Vista → [instructions for using VISTA](#)
rVista → [instructions for using rVISTA](#)

DOWNLOAD VISTA
Go to our [software download page](#) to obtain VISTA's alignment and visualization programs.

INFORMATION about VISTA
How to cite VISTA.
[Guidelines for citations, comments](#)

Vista is an integrated computational system for global alignment and visualization, designed for comparative genomics. It allows for the visualization of long sequence alignments of DNA from two or more species with annotation information, and it was developed to locate conserved sequences in syntenic regions (Dubchak et al., 2000).

It has a clean output, allowing for easy identification of sequence similarities and differences, and is easily configurable, enabling the visualization of alignments of various lengths at different levels of resolution.

This system consists of several unified modules:

aVid
the program for global alignment of DNA sequences of arbitrary length. In addition to aligning two finished sequences, it can also handle one sequence in a non-ordered and non-oriented draft format. [Details](#).

Vista
A computational tool for comparing an arbitrary number of genomic sequences from different species. [Details](#).

Modules of VISTA:

- Program for global alignment of DNA fragments of any length
- Visualization of alignment and various sequence features for any number of species
- Evaluation and retrieval of all regions with predefined levels of conservation

Aligning large genomic regions

- Long sequences lead to memory problems
- Speed becomes an issue
- Long alignments are very sensitive to parameters
- Draft sequences present a nontrivial problem
- Accuracy is difficult to measure and to achieve

References for some existing programs:

Glass:

Domino Tiling, Gene Recognition, and Mice.

Pachter, L. *Ph.D. Thesis, MIT* (1999)

Human and Mouse Gene Structure: Comparative Analysis and Application to Exon Prediction.

Batzoglou, S., Pachter, L., Mesirov, J., Berger, B., Lander, E. *Genome Research* (2000).

MUMmer

Delcher, A.L., Kasif S., Fleischmann, R.D., Peterson J., White, O. and Salzberg, S.L.

Alignment of whole genomes. *Nucleic Acids Research* (1999)

PipMaker

PipMaker: A Web Server for Aligning Two Genomic DNA Sequences.

Scott Schwartz, Zheng Zhang, Kelly A. Frazer, Arian Smit, Cathy Riemer, John Bouck, Richard Gibbs, Ross Hardison, and Webb Miller. *Genome Research* (2000)

Scan2

Dbscan/Scan2: Fast alignment of mega-sequences.

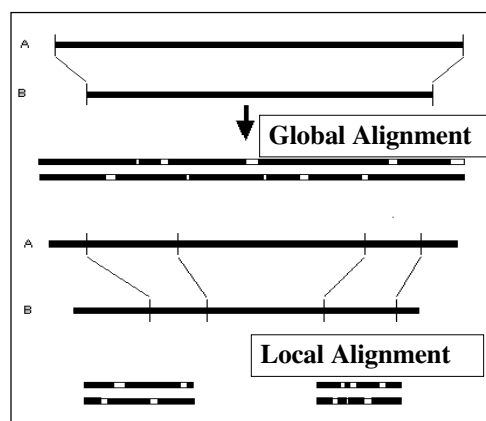
Seledtsov I.A., Solovyev V.V. To Appear. Web site <http://softberry.com/>

Local alignment algorithms are designed to search for highly similar regions in two sequences that may not be highly similar in their entirety. The algorithm works by first finding very short common segments between the input sequence and database sequences, and then expanding out the matching regions as far as possible.

For cross-species comparison one needs to accurately align two complete sequences. It is insufficient to find common similar regions in the two sequences, rather, what is needed is a global map specifying how the two sequences fit together, much like understanding how the pieces in a puzzle connect up with each other.

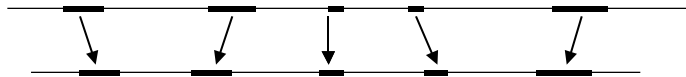
This problem is called **global alignment**

Local vs global alignment



AVID- the alignment engine behind VISTA

- Very fast global alignment of megabases of sequence.
- Provides details about ordered and oriented contigs, and accurate placement in the finished sequence.
- Full integration with repeat masking.



- ORDER and ORIENT
- FIND all common k-long words (k-mers)
- ALIGN k-mers scoring by local homology
- FIX k-mers with good local homology
- RECURSE with smaller k (shorter words)

Visualization



```
tggtacattcaaattatg-----ttctcaaagtgagcatgaca-acttttttccatgg
| | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
tgatgacatctatttgcgttttccttttagaaactgcatgagagcctggcctagtaggg
```



Window of length L is centered at a particular nucleotide in the base sequence

Percent of identical nucleotides in L positions of the alignment is calculated and plotted

Move to the next nucleotide

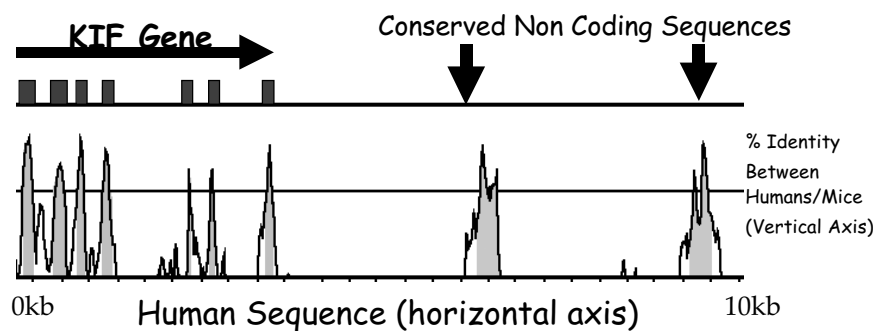
Finding conserved regions with percentage and length cutoffs

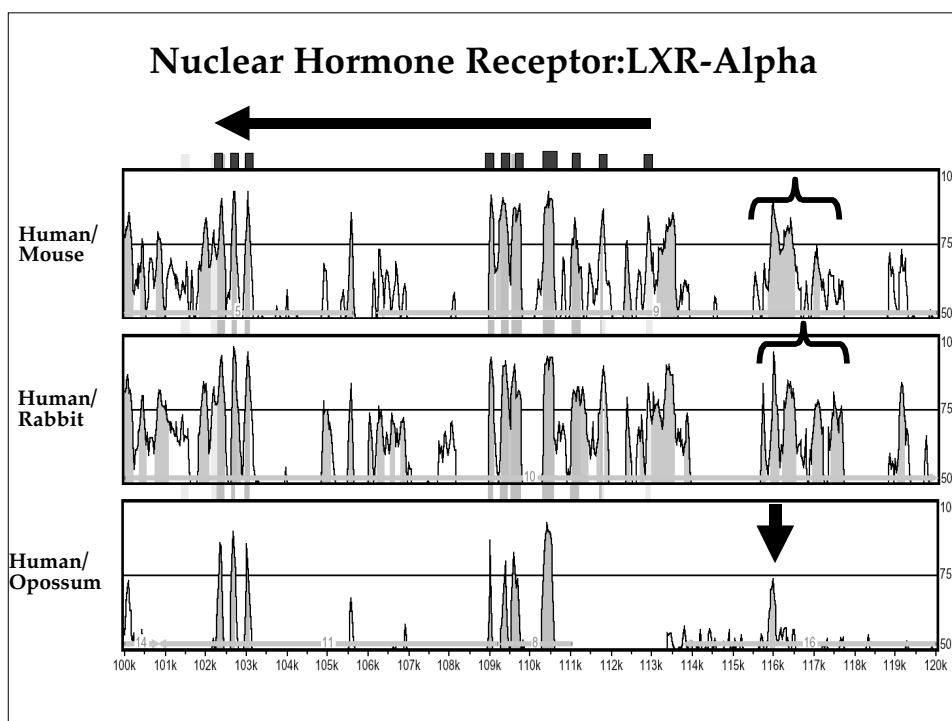
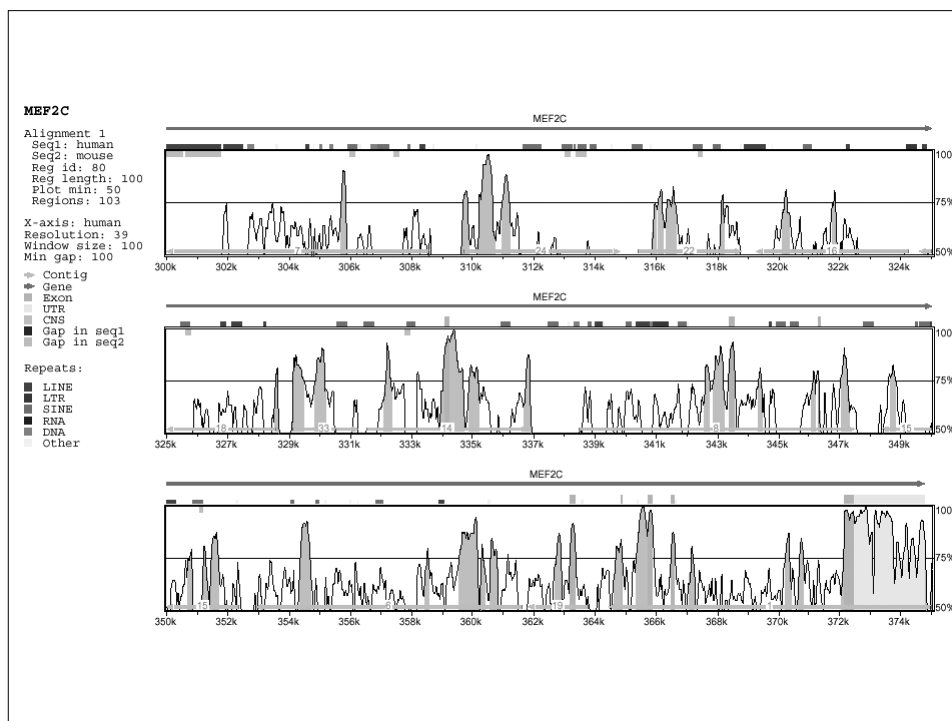
Conserved segments with percent identity X and length Y - regions in which every contiguous subsegment of length Y was at least $X\%$ identical to its paired sequence. These segments are merged to define the conserved regions.

Output:

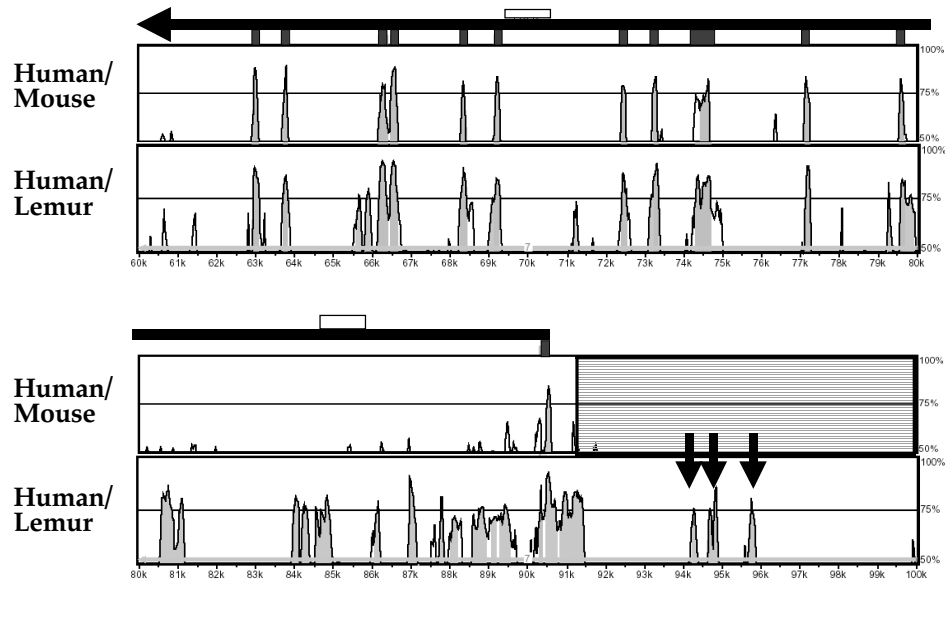
11054 - 11156 = 103bp at 77.670%	NONCODING
13241 - 13453 = 213bp at 87.793%	EXON
14698 - 14822 = 125bp at 84.800%	EXON

VISTA plot

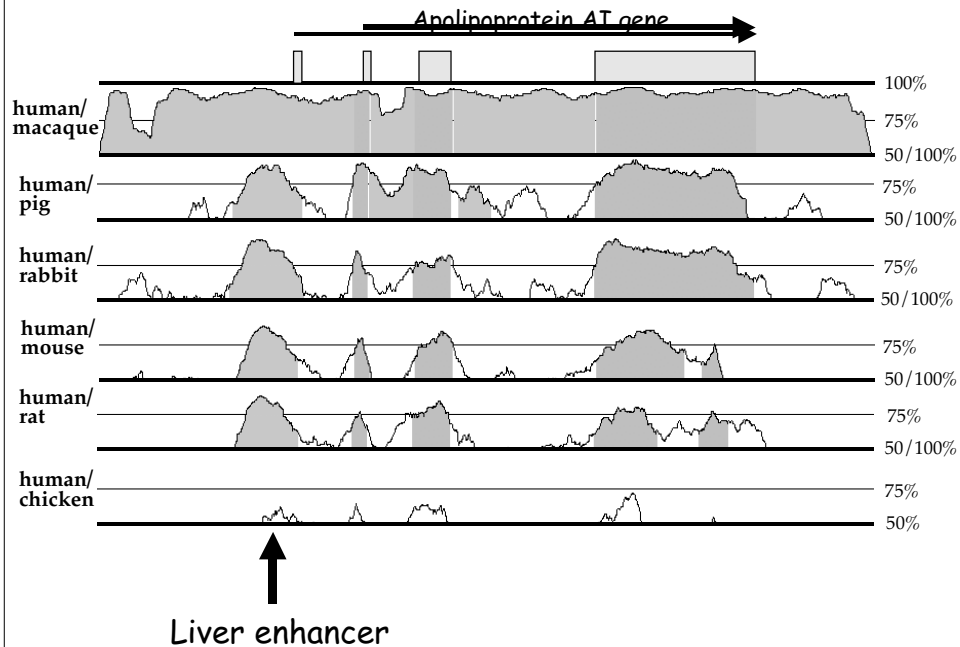




Low-Density Lipoprotein Receptor (LDLR)



Multi-Species Comparative Analysis (VISTA)



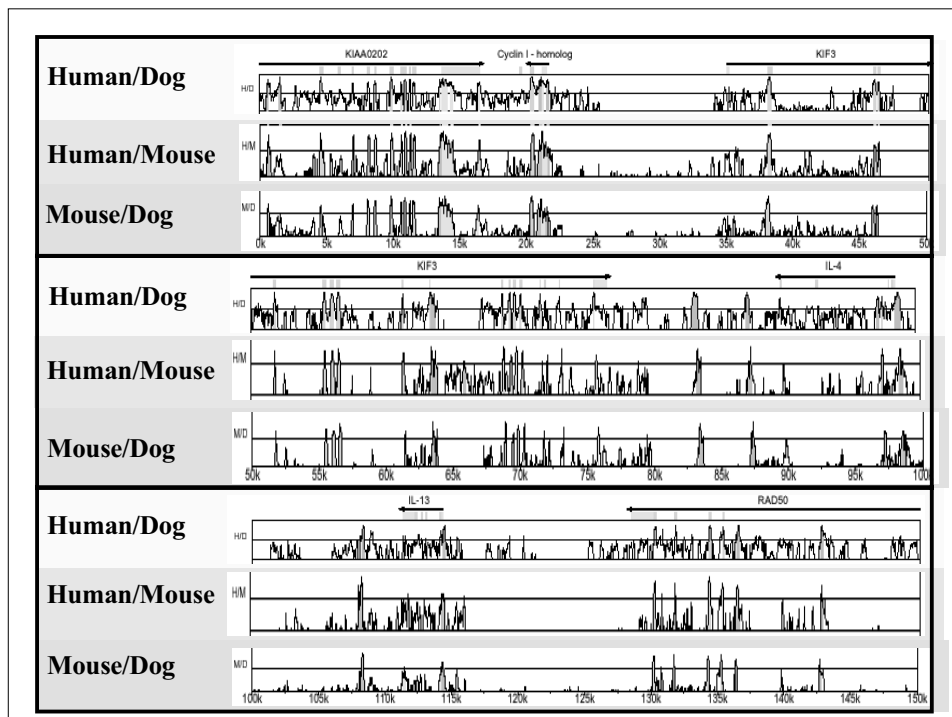
VISTA server input files

VISTA server output files

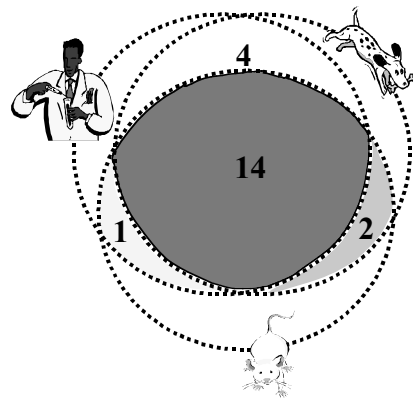
- All pair wise global alignments of the sequences
- VISTA plot
- The list of conserved regions at predefined by the user length and conservation cutoffs

VISTA flavors

- VISTA - comparing DNA of multiple organisms
- for 3 species - analyzing cutoffs to define actively conserved non-coding sequences
- cVISTA - comparing two closely related species
- rVISTA - regulatory VISTA



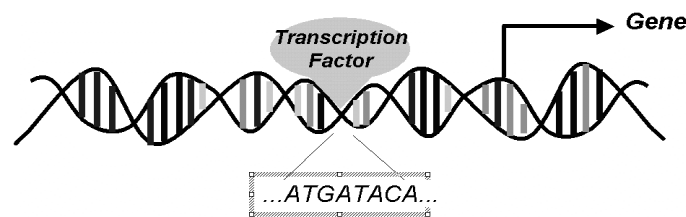
Active conservation of noncoding sequences - present in more than two mammals



% Cutoff
sum of three pair wise
Intersection/Union
values is maximal

Over 120 basepairs:
H/D > 92%
H/M > 80%
D/M > 77%

Identifying non-coding sequences (CNSs) involved in transcriptional regulation



rVISTA - prediction of transcription factor binding sites

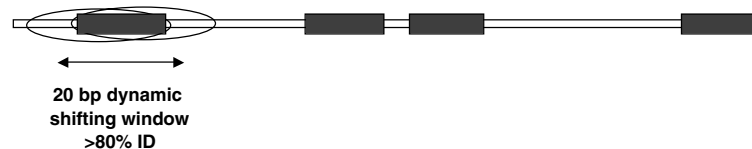
- Simultaneous searches of the major transcription factor binding site database (Transfac) and the use of global sequence alignment to sieve through the data.
- Combination of database searches with comparative sequence analysis reduces the number of predicted transcription factor binding sites by several orders of magnitude.

Regulatory VISTA (rVISTA)

1. Identify potential transcription factor binding sites for each sequence using library of matrices (TRANSFAC)
2. Identify aligned sites using VISTA
3. Identify conserved sites using dynamic shifting window

Percentage of conserved sites of the total 3-5%

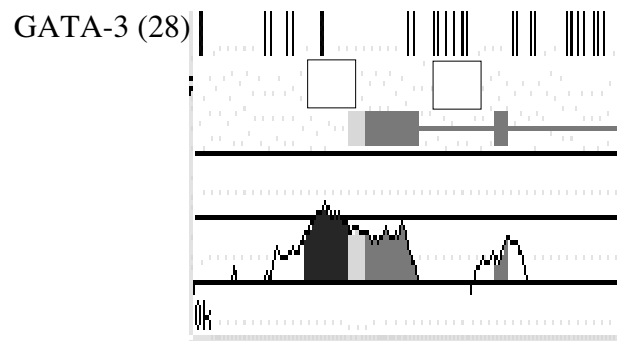
Human TGATTTCTCGGCAGCAAGGGAGGGCCCCATGACAAAGCCATTTGAAATCCCAGAAAGCAATTTCTACTTACGACCTCAGCTTTCTGTTGCTGTCTCCCTT
 Mouse TGATTTCTCGGCAGCCAGGGAGGGCCCCATGACGAAGCCACTCGAAATCCCAGAAAGCAATTTCTACTTACGACCTCAGCTTTCTGTTGCTGTCTCTCCCTT
 Dog TGATTTCTCGGCAGCAAGGGAGGGCCCCATGACGAAGCCATTTGAAATCCCAGAAAGCAATTTCTACTTACGACCTCAGCTTTCTGTTGCTGTCTCTCCCTT
 Rat TGATTTCTCGGCAGCCAGGGAGGGCCCCATGACGAAGCCACTCGAAATCCCAGAAAGCAATTTCTACTTACGACCTCAGCTTTCTGTTGCTGTCTCTCCCTT
 Cow TGATTTCTCGGCAGCCAGGGAGGGCCCCATGACGAAGCCATTTGAAATCCCAGAAAGCAATTTCTACTTACGACCTCAGCTTTCTGTTGCTGTCTCTCCCTT
 Rabbit TGATTTCTCGGCAGCCAGGGAGGGCCCCACGAC-AAGCCATTCAAATCCCAGAAAGTGAATTTCTACTTACGACCTCAGCTTTCTGTTG---CTCTCTCTCTCCCTT

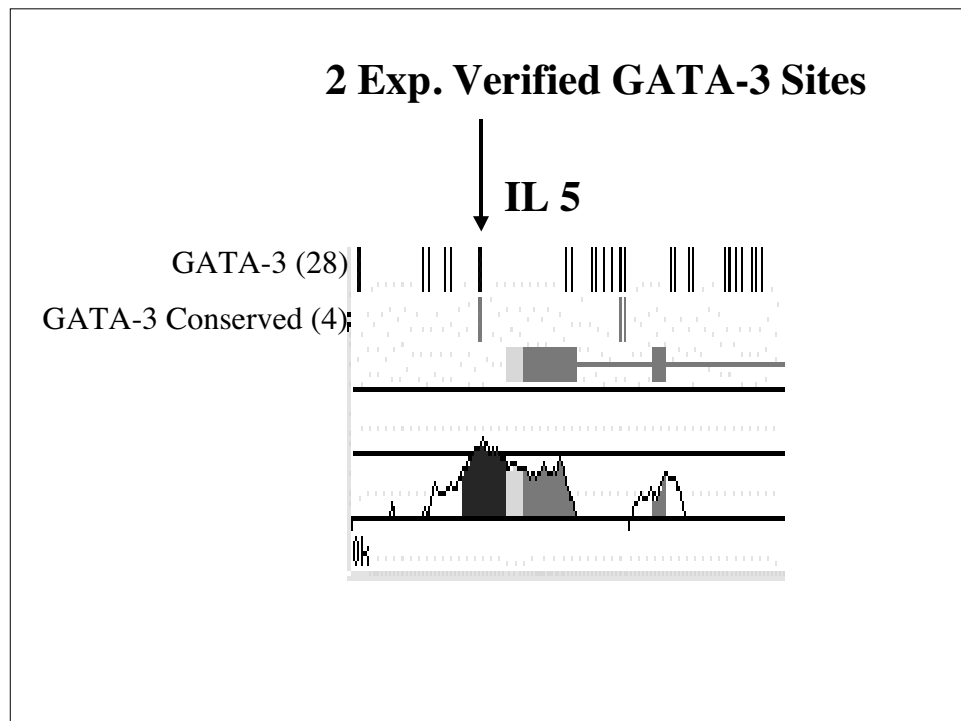
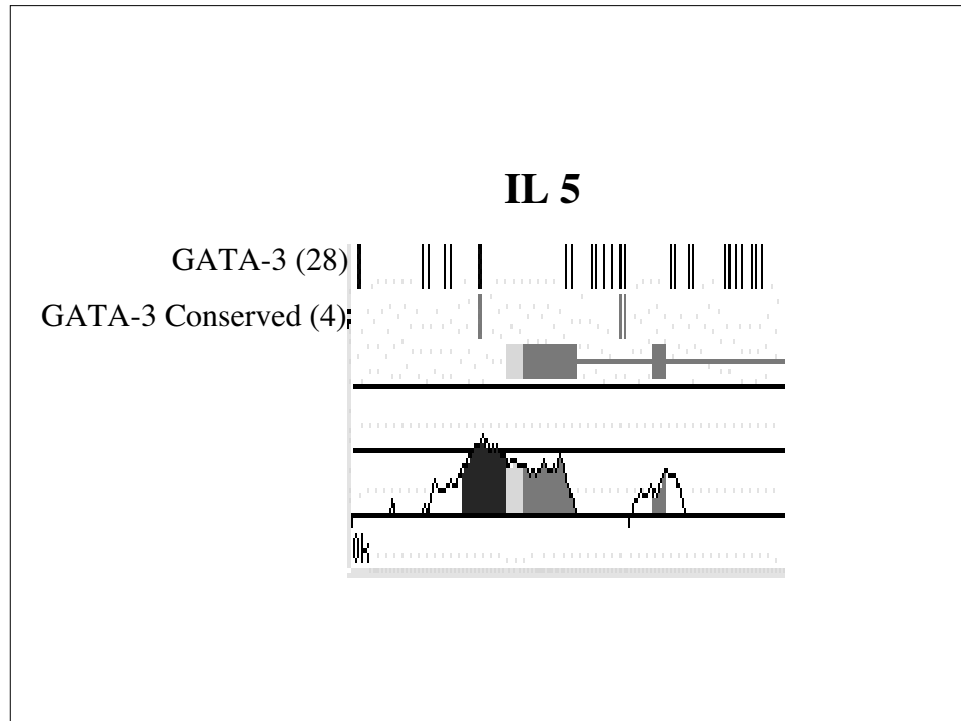


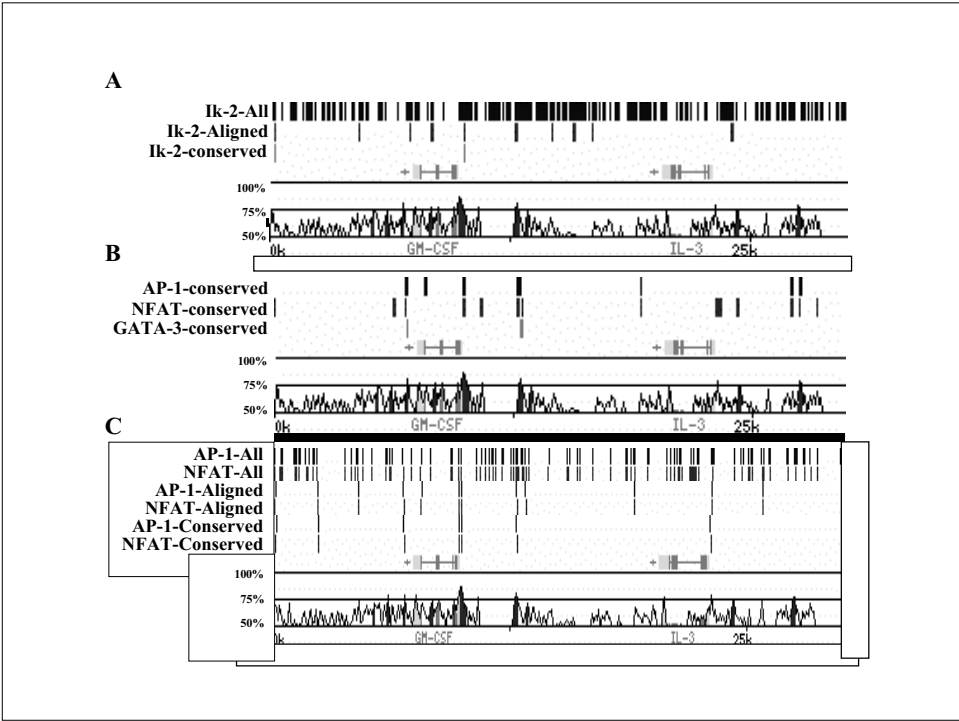
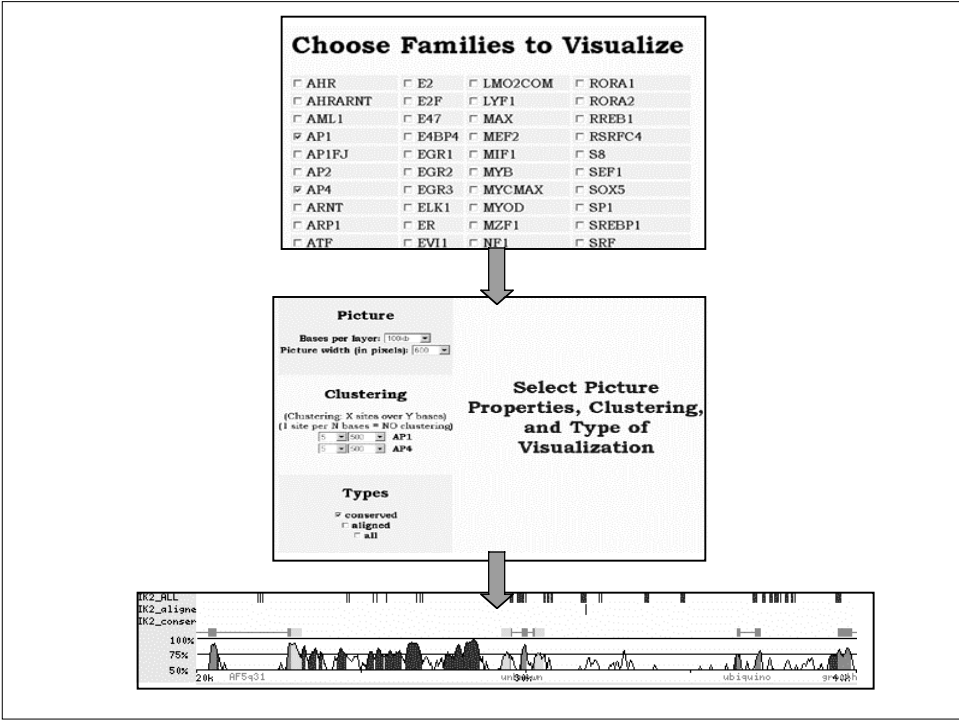
~1 Meg region, 5q31

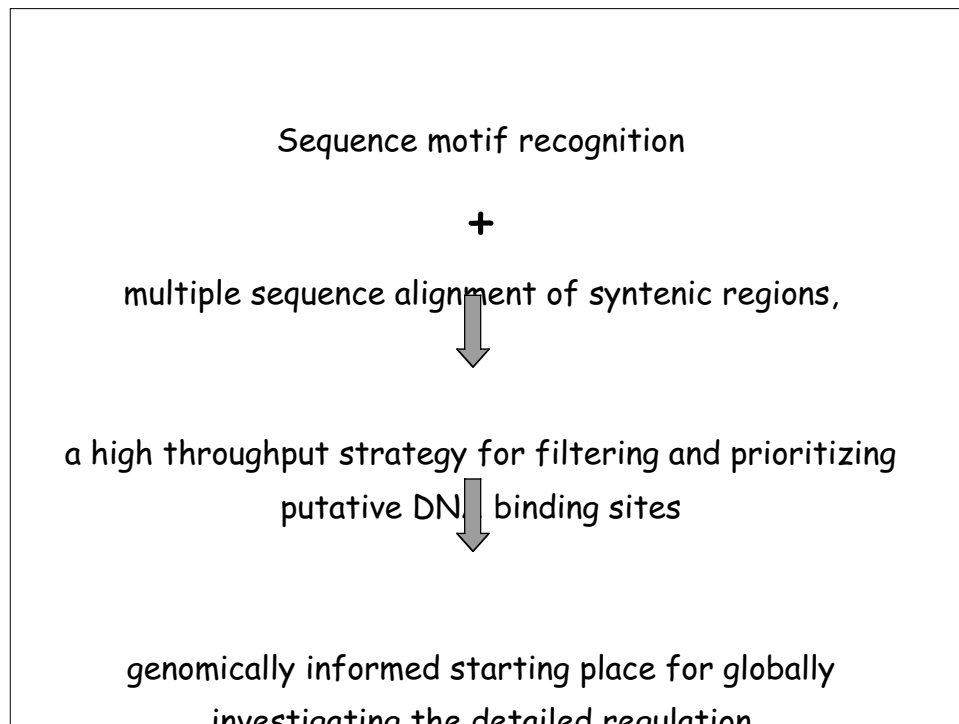
	Coding	Noncoding
Human interval Transfac predictions for <i>GATA</i> sites	839	20654
Aligned with the same predicted site in the mouse seq.	450	2618
Aligned sites conserved at 80% / 24 bp dynamic window	303	731
Random DNA sequence of the same length		29280

IL 5









Main features of VISTA

- Clear , configurable output
- Ability to visualize several global alignments on the same scale
- Alignments up to several megabases
- Working with finished and draft sequences
- Available source code and WEB site

What if you don't have a sequence of other species for the region of your interest?

Are there publicly available comparative genomics data?

Large scale VISTA applications:

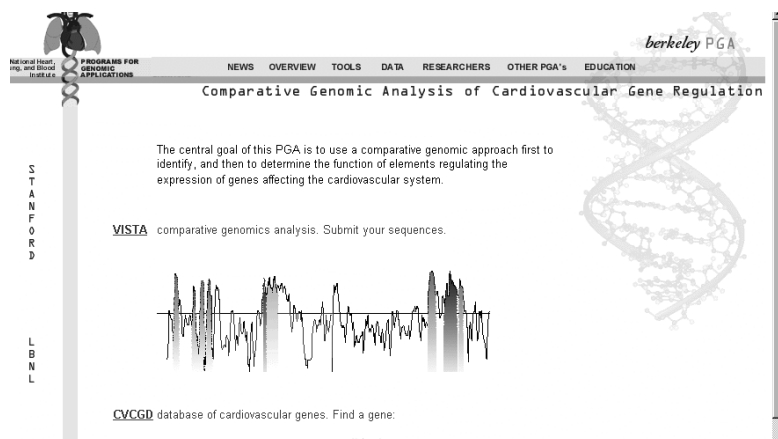
Cardiovascular comparative genomics database

<http://pga.lbl.gov>

Godzilla - comparing the human and mouse genome

<http://pipeline.lbl.gov>

<http://pga.lbl.gov>



<http://pga.lbl.gov/cvcgd.html>

berkeley PGA



Cardiovascular Comparative Genomic Database (CVCGD)

This database includes well-studied CV genes, for which an understanding of regulation should provide insights into CV relevant biological issues. While only a fraction of these genes will be characterized in the PGA biological projects over the 4-year time period of this program, the sequence of ~200 genomic intervals containing CV genes will be obtained and comparatively annotated and included in the CVCGD.

The database contains a variety of information for each gene relevant to this project:

- Gene name;
- Gene ID in the OMIM database (**OMIM**);
- Human map location (**HM**);
- GenBank accession number for human cDNA (**HC**);
- Mouse map location (**MM**);
- GenBank accession number for mouse cDNA (**MC**).

SEARCH the CVCGD

- [by gene name and abbreviation](#)
- [sorted alphabetically](#)
- [by categories](#) (groups of diseases).

Example of CVCGD entry

Solute carrier family 22, organic cation transporter member 4 (SLC22A4, OCTN1) - Netscape

File Edit View Go Communicator Help

CVCGD search results Reload Home Search Netscape Print Security Shop Stop

Location: http://pga.lbl.gov/cgi-bin/get_gene?id=234

Bioinformatics

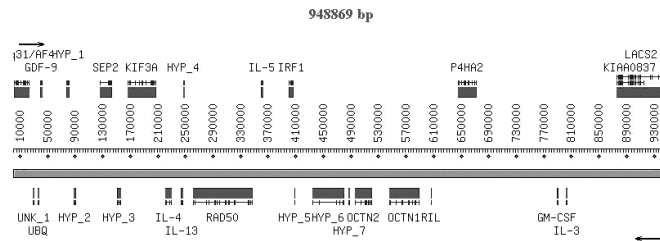
Solute carrier family 22, organic cation transporter member 4 (SLC22A4, OCTN1)

- Category: Atherosclerosis
- Gene ID in the OMIM database: [604190](#)
- Human map location: 5q31
- GenBank accession number for human cDNA: [NM_003059](#)
- Mouse map location: 11
- GenBank accession number for mouse cDNA: [NM_019687](#)
- Annotation of the human sequence
- Human mouse alignment: Whole sequence | [1-100000](#) | [100001-200000](#) | [200001-300000](#) | [300001-400000](#) | [400001-500000](#) | [500001-600000](#) | [600001-700000](#) | [700001-800000](#) | [800001-900000](#) | [900001-967696](#) (see important note below) | [Printable version \(PDF\)](#)
- [List of conserved regions](#)

Note: If your browser hangs or crashes on the alignment page you can try [this link](#) instead.

Short annotation of the region

Annotation of the VA5q31 region *

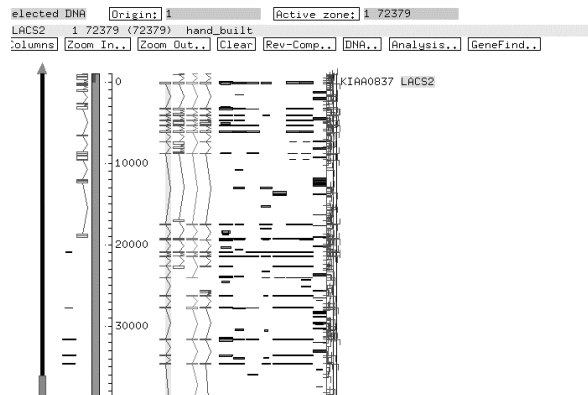


* Assembly contains a deletion of 18822 bp after the first exon of the RIL gene

Gene Name	Identity/Similarity
AF5q31/AF4	Identical to gi6601437 Homo sapiens AF5q31 protein (AF5q31) (Start not found)
GDF-9	Identical to gi488526 growth differentiation factor 9 from Homo sapiens
HYP_1	hypothetical
CEB	Identical to gi1502087 KIAA0837/Seb2 gene from Homo sapiens

Detailed annotation in AceDB format

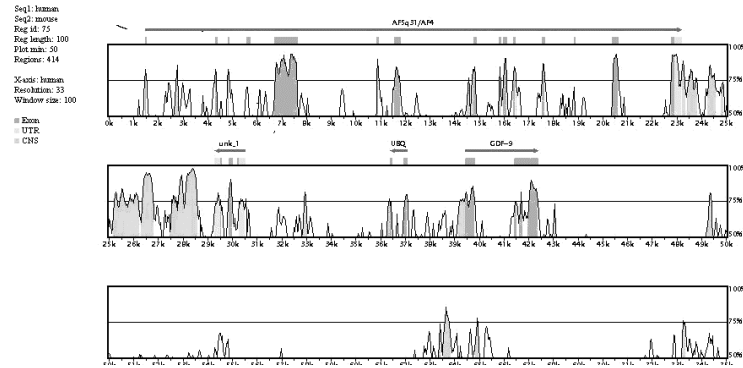
GENE LACS2 : Identical to [gi4336603](#) long-chain acyl-CoA synthetase 2 (LACS2) from Homo sapiens (End not found)



VISTA plot of the region

Genomic region containing Solute carrier family 22, organic cation transporter member 4 (SLC22A4, OCTN1)

You can view corresponding alignment regions if you click on the picture inside plot frames



Alignment

Genomic region containing Solute carrier family 22, organic cation transporter member 4 (SLC22A4, OCTN1)

seq1 = human
seq2 = mouse

	6990	7000	7010	7020	7030	7040
seq1	CAGAGTGACAGCACAAACACAGAGAAGAACTGTAGGCAAAAAACACCCAAAAAGGCTGAG					
seq2	CAGAGCGACAGTACCACCTCAGAGGAGAACTGTCTGGCAAAAAACACCCAAAAACCTGAG					
	8130	8140	8150	8160	8170	8180
	7050	7060	7070	7080	7090	7100
seq1	AAGGCAGCTGCTGAAGAGCCTCGTGGAGGCCTGAAGATAGAAAAGTGAAACCCCTGTAGAC					
seq2	AAGTCAGCTGCTGAAGAGCCTCGTGGAGGCCTGAAGATAGAAAAGTGAGACCCCTGTGGAC					
	8190	8200	8210	8220	8230	8240
	7110	7120	7130	7140	7150	7160
seq1	TTGGCTAGCAGCATGCCCTCCAGCAGACACAAAGCAGCCACCAAGGGCTCAAGGAAACCC					
seq2	ATGGCTGCCAGCATGCCCTCCAGCAGGCACAAAGCAGCCACCAAGGGCTCGAGGAAACCC					
	8250	8260	8270	8280	8290	8300

Conserved regions

Genomic region containing Solute carrier family 22, organic cation transporter member 4 (SLC22A4, OCTN1)

Criteria: 75% identity over 100 bp

***** Conserved Regions - human (mouse) *****

1469	(580)	to	1515	(626)	=	47bp	at	85.1%	exon
2668	(2043)	to	2817	(2191)	=	153bp	at	80.4%	noncoding
4316	(4531)	to	4370	(4585)	=	55bp	at	100.0%	exon
4816	(6136)	to	4853	(6173)	=	38bp	at	97.4%	exon
6717	(7860)	to	7634	(8777)	=	918bp	at	87.8%	exon
10839	(10749)	to	10927	(10837)	=	89bp	at	91.0%	exon
11553	(12627)	to	11793	(12873)	=	247bp	at	81.8%	exon
14508	(15706)	to	14622	(15823)	=	119bp	at	76.5%	noncoding
14671	(15886)	to	14783	(16003)	=	118bp	at	74.6%	noncoding
14784	(16004)	to	14878	(16098)	=	95bp	at	89.5%	exon
15797	(17526)	to	15860	(17589)	=	64bp	at	93.8%	exon
15975	(17703)	to	16111	(17839)	=	137bp	at	90.5%	exon
16365	(18045)	to	16436	(18116)	=	72bp	at	91.7%	exon
16437	(18117)	to	16535	(18217)	=	101bp	at	75.2%	noncoding
17554	(18914)	to	17647	(19007)	=	94bp	at	87.2%	exon

GODZILLA THE BERKELEY GENOME PIPELINE

Comparing the human and mouse genomes (~3 billion bases each)

Human Genome - GoldenPath Assembly at UCSC

good coverage, large contigs, many annotations
stable assembly.

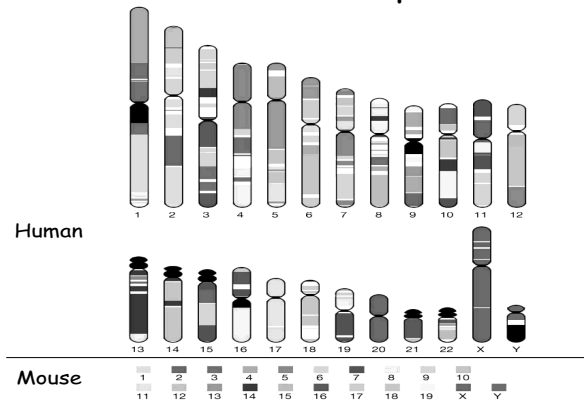
Mouse Genome:

All finished contigs from GenBank

First draft assembly ~3X of WGS, Arachne and Phusion

~5X released a week ago, at work
<http://pipeline.lbl.gov>

Chromosome Comparison



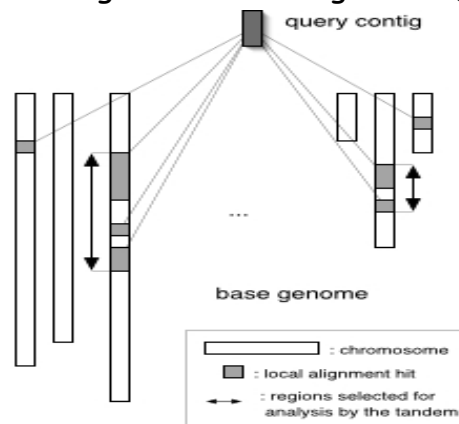
Base pair alignment

```

247 GGTGAGGTCGAGGACCCTGCA  CGGAGCTGTATGGAGGGCA  AGAGC
    |:  ||  ||||:  ||||  -:||  |||  |:|  |||---|||
368 GAGTCGGGGGAGGGGGCTGCTGTTGGCTCTGGACAGCTTGCATTGAGAGG
  
```

A Whole Genome Alignment Strategy

Step 1: Rough alignment using fixed matches to localize a contig on the human genome (BLAT)



note: proportions are not respected

Step 2: Refined global alignment of anchored regions AVID

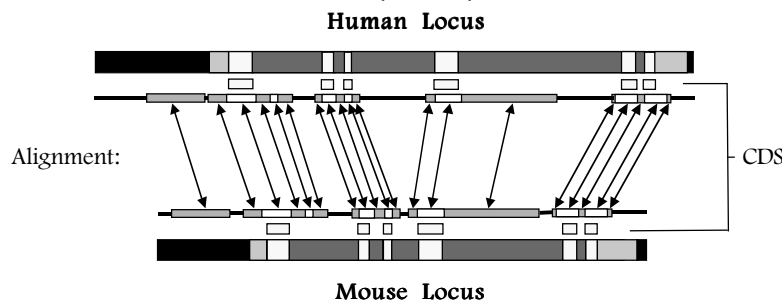
Input: Region of a few hundred kilobases long

Output: Base pair alignment

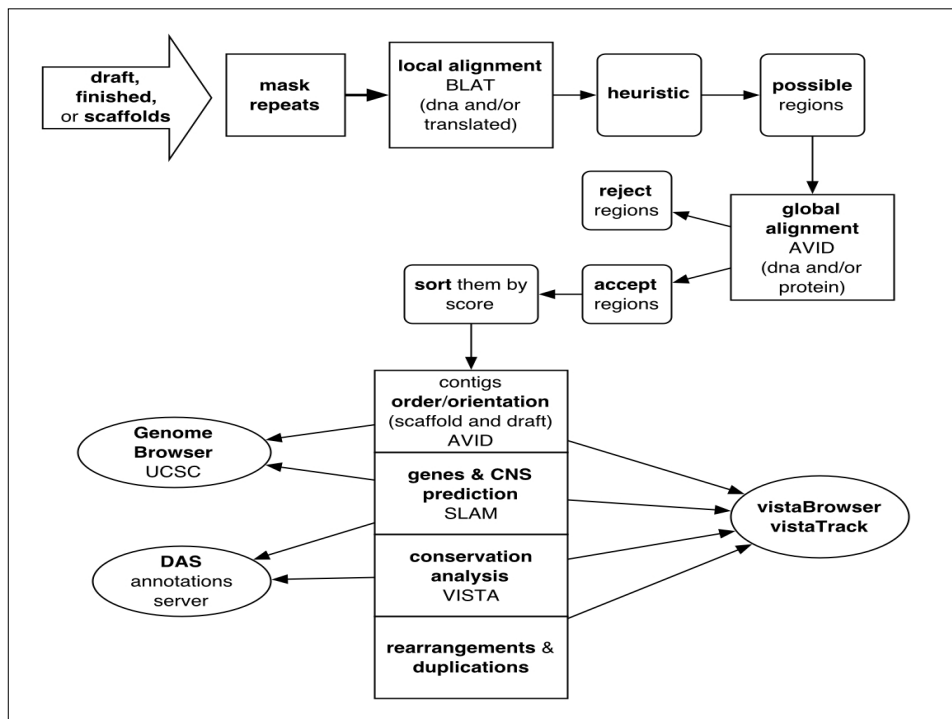
Step 3: Expensive super-fine alignment SLAM

Input: Region of a few hundred kilobases long.

Output: An alignment and annotation of conserved elements (HMM)



Step 4: Visualization



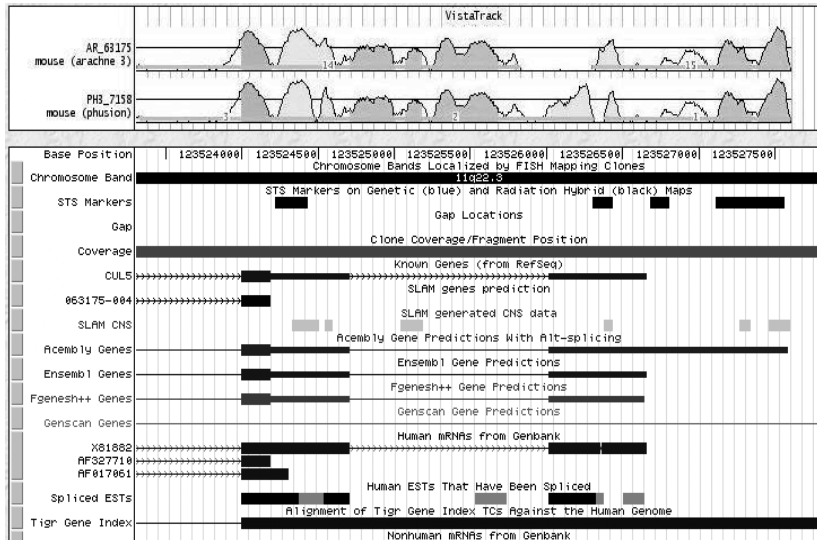
Visualization of the whole genome comparative analysis

- VISTA pictures/VISTABrowser
 - Stand-alone Java applet for detailed comparison
- VISTA Track on the *Genome Browser* from UCSC
 - Comparison in the context of the human genome annotation
- Distributed Annotation System (DAS)

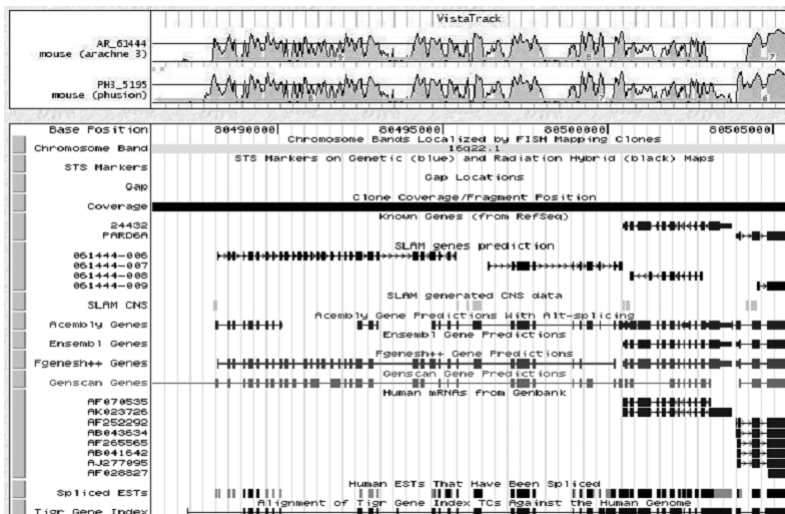
Examples of Results

- Understanding the structure of conservation
- Identification of putative functional sites
- Discovery of new genes
- Detection of contamination and misassemblies

Two mouse assemblies are better than one



New genes?



Summary

Suite of comparative genomics tools VISTA

<http://www-gsd.lbl.gov>

Godzilla comparing the human and mouse genome

<http://pipeline.lbl.gov>

Cardiovascular comparative genomics database

<http://pga.lbl.gov>

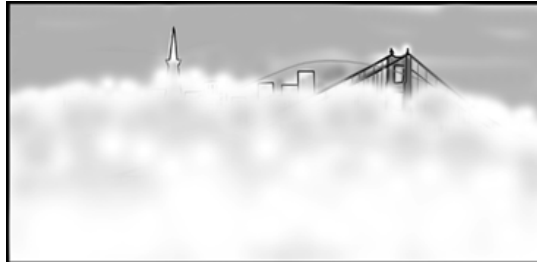
Questions? Write to vista@lbl.gov

Publications on the tools:

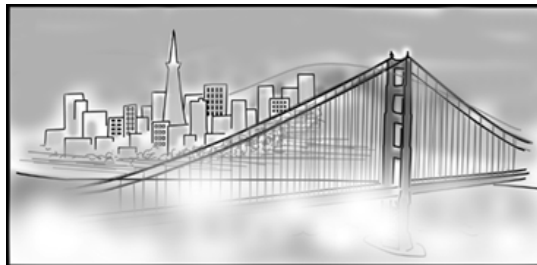
- I. Dubchak, M. Brudno, L.S. Pachter, G.G. Loots, C. Mayor, E. M. Rubin, K. A. Frazer. (2000) Active conservation of noncoding sequences revealed by 3-way species comparisons. *Genome Research*, 10: 1304-1306.
- C. Mayor, M. Brudno, J. R. Schwartz, A. Poliakov, E. M. Rubin, K. A. Frazer, Lior S. Pachter, I. Dubchak. (2000) VISTA: Visualizing global DNA sequence alignments of arbitrary length. *Bioinformatics*, 16: 1046-1047.
- G. G. Loots, I. Ovcharenko, L. Pachter, I. Dubchak and E. M. Rubin. (2002) Comparative sequence-based approach to high-throughput discovery of functional regulatory elements. *Genome Res.*, to appear
- I. Dubchak, L. Pachter. (2002) The computational challenges of applying comparative-based computational methods to whole genomes. *Briefings in Bioinformatics*, 3, 18.

Towards Better VISTAs

Information
from a Single
Sequence
Alone



Multi-Organism
High Quality
Sequences



Thanks

Biology

Kelly Frazer
Gaby Loots
Len Pennacchio

Bioinformatics

Michael Brudno
Olivier Couronne
Chris Mayor
Ivan Ovcharenko
Alexander Poliakov
Jody Schwartz

Eddy Rubin

Lior Pachter (UCB)